職業・産業自動コーディングシステムのWeb公開に向けて

高橋　和子（敬愛大学）魏　大比（東京工業大学）　田辺　俊介（東京大学）吉田　崇（東京大学）

研究の背景

- 社会調査において自由記述によって職業・産業情報を得た場合、コーディング作業が必要になる。
- 職業・産業データのコード化作業に伴う負担の軽減を目的に種々のコーダ支援用の自動コーディングシステムを開発してきたが、利用面での問題がある。
- これらのシステムを誰もが自由に利用できるようにするためにはWeb公開が有効である。

職業・産業自動コーディングシステムのWeb公開

Web画面を通じて職業・産業についての自由記述が入力されたデータファイルを送信

↓

職業・産業コードおよび確信度が付与された結果ファイルを受信

＊メンテナンスの点からソフトウェアを公開する方法は採用しない

送信ファイル例（職業コードの場合）

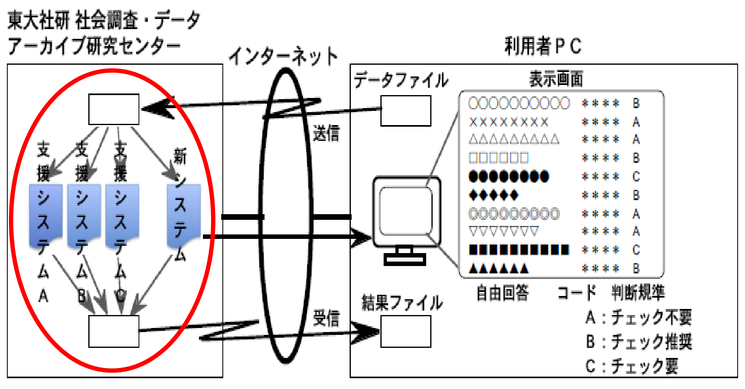
ID	学歴	地位&役職	規模	仕事の内容
1	9	9	8	工場でコピー機のトナーカートリッジの製造
2	9	3	6	工場でガラス吹き
3	11	4	9	福祉事務所で生活保護業務の現業員
4	11	8	8	予備校での事務
5	10	2	4	看護師

プリコード　　自由記述（テキスト・データ）

ID	SSM職業コード1	確信度	SSM職業コード2	確信度
1	630	C	644	C
2	625	B	689	B
3	554	B	629	C
4	554	A	560	C
5	578	A	688	C

第1候補　　第2候補

（注）A～Cは確信度
A: 信頼性高い～C: 信頼性低い



支援システムA　：
ROCCO（ルールベース手法に基づいた手法）
支援システムB　：
機械学習（SVM）による手法
支援システムC　：
機械学習（SVM）とROCCOの併合による手法

処理の流れ

順番	処理内容	処理方法（従来）	処理方法（提案）
1	前処理（文字変換など）	人手	Java一本化
2	形態素解析	Juman	
3	ROCCO	LISP プログラム	
4	出力形式の変換（EXCEL形式）	BASIC プログラム	
5	後処理（ファイルの合併など）	人手	

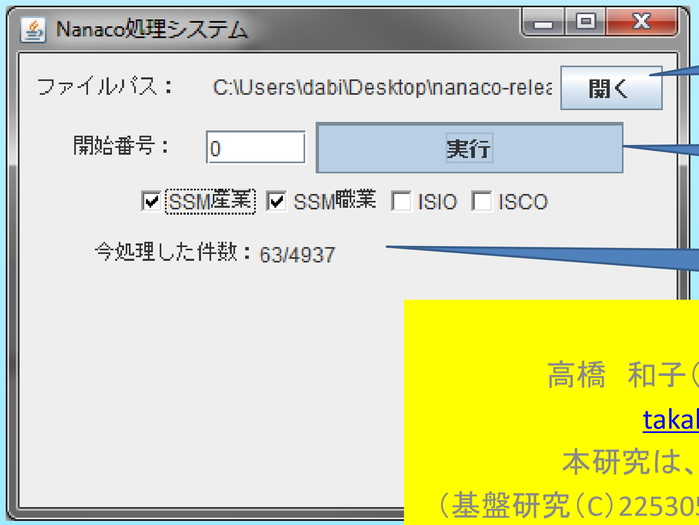
juman*　：　京都大学長尾研究室開発　形態素解析用フリーソフト

今後の予定

［2011年度］
- 東大社研附属社会調査・データアーカイブ研究センターHP画面の更新
- 機械学習手法による処理の自動化
- 機械学習とルールベース手法の統合手法による処理の自動化

［2012年度］
- 自動コーディング結果に対する確信度付与（新システム）の開発と処理の自動化など

（例）ROCCOによるSSM職業・産業コードの付与



連絡先
高橋　和子（敬愛大学国際学部）
takak@u-keiai.ac.jp

本研究は、平成22年度科研費（基盤研究（C）22530516）による成果の一部である